Biometrical Journal **44** (2002) 5, 558–570

Testing the Intraclass Version of Kappa Coeffcient of Agreement with Binary Scale and Sample Size Determination

JUN-MO NAM

Biostatistics Branch, DCEG
National Cancer Institute
Rockville, Maryland
U.S.A.

Summary

The intraclass version of kappa coefficient has been commonly applied as a measure of agreement for two ratings per subject with binary outcome in reliability studies. We present an efficient statistic for testing the strength of kappa agreement using likelihood scores, and derive asymptotic power and sample size formula. Exact evaluation shows that the score test is generally conservative and more powerful than a method based on a chi-square goodness-of-fit statistic (DONNER and ELIASZIW, 1992, *Statistics in Medicine* **11**, 1511–1519). In particular, when the research question is one directional, the one-sided score test is substantially more powerful and the reduction in sample size is appreciable.

Key words: Exact evaluation; Kappa coefficient; Power; Sample size; Score test.

1. Introduction

The kappa coefficient has been widely accepted as a measure of reliability between two ratings per subject on a binary scale. There are two types of kappa, SCOTT's index (1955) and COHEN's kappa (1960). The former is based on a model that the probability of positive classification by the 1st rating and that by the 2nd ratings are the same while the latter assumes that two probabilities are different. In this paper, we consider the former which is the intraclass version of the kappa or intra-rater correlation coefficient. Scott's kappa has also been discussed by BLOCK and KRAMER (1989) and DUNN (1989). It is interesting to note that the kappa is algebraically equivalent to the inbreeding coefficient (WRIGHT, 1951) in population genetics. Recently, DONNER and ELIASZIW (1992) have proposed a statistical method in which testing significance, power and sample size were obtained using a chi-square goodness-of-fit procedure. The method is based on a two-sided test. When researchers are interested in the strength of agreement and test whether the kappa coefficient is at least greater than a reasonable value, then a one-sided test procedure is appropriate. In this paper, we develop an efficient statistical method for inference about the coefficient of agreement, the related power and sample size

 0323-3847/02/0507-0558 $ 17.50+.50/0

determination. In Section 2, we define notation and model, and derive the likelihood score test statistic. In Section 3, asymptotic power functions of one-sided and two-sided score tests are presented and they are compared with the goodness-of-fit procedure in terms of asymptotic and exact power. In Section 4, we provide approximate sample size formulae for designing a study using the score methods. Sections 5 and 6 give an example and concluding remarks.

2. Test Statistic

Suppose that a single examiner rates each of n subjects twice, the 2^{nd} rating without recollection of the 1^{st} one as either positive (+) or negative (−). Alternatively, suppose that two examiners having the same probability of a positive classification of a subject rate n subjects independently. In either case, the n pairs of ratings can be divided into three categories: (+, +); (+, −) or (−, +); and (−, −). The observed numbers of pairs in the categories are x_2, x_1 and x_0 and their corresponding probabilities are P_2, P_1 and P_0 where the subscripts represent the number of positive ratings in a pair. Denote the probabilities of a positive and a negative rating as $Pr\,(+) = p$ and $Pr\,(-) = q$, respectively. Define the kappa, κ, as the correlation coefficient between the two ratings in a pair, i.e., $\kappa = (P_2 - p^2)/(pq) = (P_0 - q^2)/(pq)$, and consider the following multinomial model: $P_2(\kappa, p) = p^2 + pq\kappa$, $P_1(\kappa, p) = 2pq(1-\kappa)$ and $P_0(\kappa, p) = q^2 + pq\kappa$ (MAK, 1988; BLOCH and KRAMER, 1989). The log-likelihood is the expressed as

$$\ln L(\kappa, p) = x_2 \cdot \ln\{p(p + q\kappa)\} + x_1 \cdot \ln\{2pq(1-\kappa)\} + x_0 \cdot \ln\{q(q + p\kappa)\}, \quad (1)$$

where $q = 1 - p$. Kappa is the parameter of interest and p is a nuisance parameter. The maximum likelihood estimators (MLE's) of κ and p are $\hat{\kappa} = (4x_0x_2 - x_1^2)/\{(2x_0 + x_1)(2x_2 + x_1)\}$ and $\hat{p} = (2x_2 + x_1)/(2n)$, respectively. The asymptotic variance of $\hat{\kappa}$ is $\mathrm{var}(\hat{\kappa}) = (1-\kappa)\{(1-\kappa)(1-2\kappa) + \kappa(2-\kappa)/(2pq)\}/n$ (e.g., BLOCH and KRAMER, 1989; HALE and FLEISS, 1993). The first-order partial derivatives of the log-likelihood (1) are

$$S_\kappa(\kappa, p) \equiv \partial \ln L/\partial\kappa = \{x_2/(p + q\kappa) + x_0/(q + p\kappa) - n\}/(1-\kappa).$$

$S_p(\kappa, p) \equiv \partial \ln L/\partial p = A_1/(pq) + A_2/\{(p + q\kappa)(q + p\kappa)\}$ where $A_1 = x_2 + x_1 - (n + x_1)p$ and $A_2 = \{x_2 - x_0\kappa - (x_0 + x_2)(1-\kappa)p\}(1-\kappa)$. Consider a one sided score test for null-hypothesis H_0: $\kappa = \kappa_0$ against the alternative H_1: $\kappa = \kappa_1 (> \kappa_0)$. The MLE of p for a given $\kappa = \kappa_0$ is obtained from $S_p(\kappa_0, \tilde{p}) = 0$ which leads to the solution of a cubic equation:

$$a_0\tilde{p}^3 + a_1\tilde{p}^2 + a_2\tilde{p} + a_3 = 0,$$

where

$$a_0 = 2n(1-\kappa_0)^2,\ a_1 = -\{3n(1-\kappa_0)+x_2-x_0\}\ (1-\kappa_0),$$
$$a_2 = 2x_2 + x_1 - 2(2n-x_0)\ \kappa_0 + n\kappa_0^2$$

and

$$a_3 = (x_1+x_2)\ \kappa_0 .$$

Denoting $b_i = a_i/a_0$ for $i = 1, 2$ and 3, $c_1 = b_2 - b_1^2/3$ and $c_2 = b_3 - b_1 b_2/3 + 2(b_1/3)^3$, we have the MLE as

$$\tilde{p} = -2(-c_1/3)^{1/2} \cdot \cos(\pi/3 + \theta/3) - b_1/3 , \tag{2}$$

where

$$\cos\theta = (27)^{1/2} \cdot c_2/\{2c_1(-c_1)^{1/2}\}$$

by using a trigonometric method (USPENSKY, 1948). Let the score evaluated at $\kappa = \kappa_0$ and $p = \tilde{p}$ be $S_\kappa(\kappa_0, \tilde{p})$. The estimated asymptotic variance of the score evaluated at $\kappa = \kappa_0$ and $p = \tilde{p}$ is $2n\tilde{p}\tilde{q}/A_3$, where

$$A_3 = \{2\tilde{p}\tilde{q}(1-\kappa_0)\ (1-2\kappa_0) + \kappa_0(2-\kappa_0)\}\ (1-\kappa_0) \text{ and } q = 1-\tilde{p}.$$

Using the theory of BARTLETT (1953), the score statistic for testing $\kappa = \kappa_0$ against $\kappa = \kappa_1 (> \kappa_0)$ is explicitly expressed as

$$z_s = \{S_\kappa(\kappa_0, \tilde{p})\}/(2n\tilde{p}\tilde{q}/A_3)^{1/2} , \tag{3}$$

where $S_\kappa(\kappa_0, \tilde{p}) = \{x_2/(\tilde{p} + \tilde{q}\kappa_0) + x_0/(\tilde{q} + \tilde{p}\kappa_0) - n\}/(1-\kappa_0)$. Since the $2n\tilde{p}\tilde{q}/A_3$ is a consistent estimator of the asymptotic variance of $S_\kappa\{\kappa_0, \tilde{p}\}$, z_s is asymptotically distributed as normal with mean zero and variance one under $\kappa = \kappa_0$. For the one-sided score test, we have asymptotically $E_1(z_s) > E_0(z_s) = 0$ where $\kappa_1 > \kappa_0$. We reject H_0 in favor of H_1 at level α if $z_s \geq z_{(1-\alpha)}$ where $z_{(1-\alpha)}$ is the $100 \times (1-\alpha)$ percentile point of the standardized normal distribution and we do not reject H_0 otherwise. For a two-sided score test for H_0: $\kappa = \kappa_0$ against $H_1 : \kappa \neq \kappa_0$, we reject H_0 at α level if $z_s^2 \geq z_{(1-\alpha/2)}^2$. For the special case of $\kappa = 0$, we have $\tilde{p} = (2x_2 + x_1)/(2n)$ from $S_p(0, \tilde{p}) = 0$. From (3), the score test for $\kappa_0 = 0$ against $\kappa_1 (> 0)$ is $z_s = \{n^{1/2}(4x_2x_0 - x_1^2)\}/\{(2x_2 + x_1)\ (2x_0 + x_1)\}$. The square of z_s is identical to the statistic for testing the Hardy-Weinberg law ($F = 0$) by, e.g., SMITH (1970). DONNER and ELIASZIW (1992) have suggested a goodness-of-fit (GOF) statistic for testing $\kappa = \kappa_0$ against $\kappa = \kappa_1 (> \kappa_0)$ as

$$X^2 = \sum_{i=0}^{2} \{x_i - n \cdot P_i(\kappa_0, \hat{p})\}^2/\{n \cdot P(\kappa_0, \hat{p})\} , \tag{4}$$

where $P_2(\kappa_0, \hat{p}) = \hat{p}^2 + \hat{p}\hat{q}\kappa_0$, $P_1(\kappa_0, \hat{p}) = 2\hat{p}\hat{q}(1-\kappa_0)$ and $P_0(\kappa_0, \hat{p}) = \hat{q}^2 + 2\hat{p}\hat{q}\kappa_0$ with $\hat{p} = (2x_2 + x_1)/(2n)$ and $\hat{q} = 1 - \hat{p}$. The statistic (4) is asymptotically a chi-square with one degree of freedom under H_0. The GOF test is two-sided. The

square of the score test is different from the GOF procedure except for the case of testing $\kappa_0 = 0$.

3. Power of Test

Consider the power of the score test and the GOF method of the strength of kappa agreement.

3.1 *Asymptotic power*

From the score evaluated at $\kappa = \kappa_0$ and $p = \tilde{p}$, we have the expected score under H_1: $\kappa = \kappa_1$ as

$$E_1\{S_\kappa(\kappa_0, \tilde{p})\} = n(e_1 + e_2 - 1)/(1 - \kappa_0) \equiv n \cdot D\,, \tag{5}$$

where $e_1 = p(p + q\kappa_1)/(\bar{p} + \bar{q}\kappa_0)$, $e_2 = q(q + p\kappa_1)/(\bar{q} + \bar{p}\kappa_0)$, and $\bar{p}$ and $\bar{q}$ are the asymptotic limits, for large n, of $\tilde{p}$ and $\tilde{q}$ (Appendix 1). The asymptotic variances of the score under H_0: $\kappa = \kappa_0$ and H_1: $\kappa = \kappa_1$ are

$$\begin{aligned}\mathrm{var}_0\{S_\kappa(\kappa_0, \tilde{p})\} &= 2n\bar{p}\bar{q}/[\{2\bar{p}\bar{q}(1 - \kappa_0)\ (1 - 2\kappa_0) \\ &\qquad + \kappa_0(2 - \kappa_0)\}\ (1 - \kappa_0)] \equiv n \cdot v_0\,, \\ \mathrm{var}_1\{S_\kappa(\kappa_0, \tilde{p})\} &= I^*_{11} - I^{*2}_{12}/I^*_{22} \equiv n \cdot v_1\,, \end{aligned} \tag{6}$$

where I^*'s are defined in Appendix 2. The asymptotic power of the right-hand score test at level α is

$$P_r\{z_s \geq z_{(1-\alpha)} \mid H_1 : \kappa = \kappa_1 (> \kappa_0)\} = 1 - \Phi(u)\,, \tag{7}$$

where $u = \{z_{(1-\alpha)} \cdot v_0^{1/2} - n^{1/2} \cdot D\}/v_1^{1/2}$ and Φ is the cumulative standard normal distribution. The asymptotic power of the two-sided score test of size α is expressed as $P_r\{z_s^2 \geq z^2_{(1-\alpha/2)} \mid H_1 : \kappa = \kappa_1\} = 1 - \Phi(u_1) + \Phi(u_2)$ where $u_1 = \{z_{(1-\alpha/2)} \cdot v_0^{1/2} - n^{1/2} \cdot D\}/v_1^{1/2}$ and $u_2 = \{z_{(\alpha/2)} \cdot v_0^{1/2} - n^{1/2} \cdot D\}/v_1^{1/2}$. Since z_s^2 is asymptotically distributed as a non-central chi-square with one degree of freedom and a non-centrality parameter,

$$\lambda_1 = n \cdot D^2/v_0\,, \tag{8}$$

the asymptotic power of the test for given n and α can be found approximately from tables of the cumulative non-central chi-square distribution (HAYNAM, GOVINDARAJULA, and LEONE, 1970). Under H_1, the GOF statistic is also a non-central chi-square with one degree of freedom and non-centrality parameter

$$\lambda_2 = n \cdot \sum_{i=0}^{2} \{P_i(\kappa_1, p) - P_i(\kappa_0, p)\}^2/P_i(\kappa_0, p) \equiv n \cdot \Delta(\kappa_0, \kappa_1, p) \tag{9}$$

Table 1

Exact type 1 error probabilities corresponding to a nominal 0.05 level of one-sided or two-sided score tests and the chi-square goodness-of-fit method.

p	κ_0	one-sided score test	two-sided score test	GOF method	one-sided score test	two-sided score test	GOF method
			$n = 20$			$n = 50$	
.1	.1	.065	.065	.033	.072	.044	.025
	.2	.046	.046	.042	.050	.032	.030
	.3	.078	.049	.049	.063	.042	.043
	.4	.046	.034	.017	.044	.043	.055
	.5	.053	.030	.043	.045	.049	.065
	.6	.035	.020	.057	.047	.048	.066
	.7	.005	.032	.080	.040	.051	.053
.3	.1	.036	.047	.047	.046	.052	.052
	.2	.038	.052	.058	.043	.048	.052
	.3	.037	.051	.051	.047	.050	.054
	.4	.047	.044	.061	.043	.054	.055
	.5	.034	.055	.056	.042	.048	.052
	.6	.033	.051	.055	.038	.053	.054
	.7	.045	.048	.049	.035	.050	.055
.5	.1	.044	.045	.045	.041	.049	.049
	.2	.043	.050	.050	.047	.056	.056
	.3	.040	.050	.050	.034	.048	.048
	.4	.035	.049	.053	.039	.050	.050
	.5	.024	.065	.065	.044	.050	.050
	.6	.049	.049	.049	.044	.052	.052
	.7	.038	.042	.042	.044	.047	.047
Average		.041	.046	.050	.045	.048	.051

(MITRA, 1958; MENG, and CHAPMAN, 1966; DONNER and ELIASZIW, 1992). The asymptotic powers involving (8) and (9) are derived by assuming that H_1 is in a neighborhood of H_0 for a large sample size.

Numerical calculations of asymptotic powers of the three tests for various values of n, p, κ_0 and κ_1 in Table 2 show that the one-sided score test is most powerful as expected and the two-sided score test has better power than the GOF procedure.

3.2 *Exact power evaluation*

The formulae for asymptotic power in Subsection 3.1 are applicable for a large sample size. However, the available sample size of a reliability study in a typical situation may not be large. In order to complete a power comparison, it is necessary to have an exact power evaluation. Such an evaluation, also, enables us to

Table 2
Exact and asymptotic power of one-sided score test, the two-sided score test and the chi-square goodness-of-fit method for $\alpha = 0.05$ (Those in parentheses are asymptotic values).

p	κ_0	κ_1	one-sided score test	two-sided score test	GOF method	one-sided score test	two-sided score test	GOF method
				$n = 20$			$n = 50$	
.1	.2	.4	.149(.203)	.148(.136)	.133(.099)	.259(.306)	.197(.219)	.182(.176)
		.6	.338(.424)	.335(.337)	.312(.254)	.601(.652)	.542(.564)	.507(.535)
		.8	.601(.631)	.600(.554)	.583(.493)	.886(.872)	.870(.825)	.832(.867)
	.4	.6	.141(.161)	.104(.098)	.039(.096)	.232(.264)	.189(.174)	.136(.166)
		.8	.328(.385)	.274(.278)	.122(.238)	.635(.667)	.591(.554)	.478(.504)
	.6	.8	.115(.136)	.013(.072)	.010(.105)	.281(.264)	.176(.160)	.143(.192)
.3	.2	.4	.180(.221)	.126(.140)	.115(.140)	.366(.396)	.262(.282)	.263(.279)
		.6	.503(.546)	.407(.423)	.377(.412)	.875(.871)	.797(.793)	.797(.784)
		.8	.869(.844)	.822(.758)	.786(.740)	.998(.994)	.995(.987)	.995(.984)
	.4	.6	.219(.216)	.105(.130)	.086(.148)	.403(.414)	.286(.289)	.281(.300)
		.8	.634(.595)	.433(.450)	.372(.443)	.943(.932)	.892(.869)	.885(.819)
	.6	.8	.207(.226)	.123(.127)	.091(.177)	.485(.496)	.350(.345)	.326(.370)
.5	.2	.4	.209(.222)	.107(.137)	.107(.149)	.416(.417)	.293(.295)	.293(.303)
		.6	.581(.579)	.401(.442)	.400(.447)	.925(.913)	.863(.845)	.863(.823)
		.8	.935(.897)	.857(.816)	.856(.782)	1.000(.999)	.999(.997)	.999(.991)
	.4	.6	.202(.229)	.071(.137)	.071(.164)	.435(.455)	.302(.322)	.302(.339)
		.8	.668(.652)	.391(.496)	.387(.497)	.973(.967)	.939(.924)	.939(.870)
	.6	.8	.295(.252)	.121(.143)	.117(.201)	.591(.562)	.421(.404)	.421(.424)

examine the adequacy of calculating asymptotic power for small sample size.

The exact power of the one-sided score test (EPS) for κ_0 against $\kappa_1 (> \kappa_0)$ at level α is

$$EPS = \sum_{x \in R} Pr(\boldsymbol{x} \mid \kappa_1, p), \tag{10}$$

where R is a region of sampling points such that $z_s \geq z_{(1-\alpha)}$ and $P_r(\boldsymbol{x} \mid \kappa_1, p) = n! \prod_{i=0}^{2} \{P_i(\kappa_1, p)^{x_i}/x_i!\}$. The exact level corresponding to the nominal α of the test is a special case of (10), i.e., $\kappa_1 = \kappa_0$. Exact type 1 error probabilities and powers of the two-sided score test and the GOF method are similarly formulated.

The calculations of actual level of significance corresponding to a nominal $\alpha = 0.05$ of the three tests for various values of p, κ_0 and n are summarized in Table 1. It shows that exact type 1 error probabilities of the tests are reasonably close to the 0.05 level and the score tests are, in general, conservative when sample size is small. When a null-hypothesis, κ_0, approaches the upper limit of the kappa coefficient, i.e., $\kappa_0 \to 1$, the right-hand sided score test becomes degenerate, i.e., an exact type 1 error probability approaches zero, for small n. In fact, all three tests are inapplicable for this extreme case.

Using (10), the exact powers of the three tests at $\alpha = 0.05$ for various configuration of values of n, p, κ_0, and κ_1 are presented in Table 2. Powers for $p = 0.7$ and 0.9 are those for $p = 0.3$ and 0.1 in Table 2, i.e., they are symmetric about $p = 0.5$. Exact power evaluation indicates that the one-sided score test is the most powerful among the three tests and the two-sided score test is more powerful than the GOF method. This is consistent with the conclusion based on asymptotic powers. Considering the conservative nature of the score method and power performance, an order of preference is apparent particularly for a small sample size, i.e., one-sided, two-sided score tests and GOF method. Numerical results in Table 2 show that asymptotic powers for small or moderate sample size are fairly close to exact ones except some extreme cases.

4. Approximate Sample Size

From the asymptotic power function (7), an approximation to the sample size required to achieve power, $1 - \beta$, for the one-sided score test at level α is obtained by using a general relation (e.g., NAM, 1995): $n \cdot D = z_{(1-\alpha)} \cdot (n\nu_0)^{1/2} + z_{(1-\beta)} \cdot (n\nu_1)^{1/2}$ where D, ν_0 and ν_1 are found from (5) and (6). The approximate sample size required for a specific power of the test is expressed as

$$n = \{z_{(1-\alpha)} \cdot \nu_0^{1/2} + z_{(1-\beta)} \cdot \nu_1^{1/2}\}^2 / D_2 \,. \tag{11}$$

The sample size is inversely related to power and also to departure of κ_1 from κ_0. From (8), an approximate sample size for power, $1 - \beta$, of the two-sided score test at α is written as

$$n = \lambda(1, 1 - \beta, \alpha) \cdot \nu_0 / D_2 \,, \tag{12}$$

where $\lambda(1, 1 - \beta, \alpha)$ is the value of the non-centrality parameter of the cumulative non-central chi-square distribution with one degree of freedom corresponding to power $1 - \beta$ and α level. DONNER and ELIASZIW (1992) provided a sample size using the GOF procedure as

$$n = \lambda(1, 1 - \beta, \alpha) / \Delta(\kappa_0, \kappa_1, p) \,, \tag{13}$$

where

$$\Delta(\kappa_0, \kappa_1, p) = pq(\kappa_1 - \kappa_0)^2 \, (\delta_1 + \delta_2 + \delta_3), \; \delta_1 = q/(p + q\kappa_0),$$
$$\delta_2 = 2/(1 - \kappa_0) \text{ and } \delta_3 = p/(q + p\kappa_0)$$

from (9). For the special case of $\kappa_0 = 0$, we have $\Delta(0, \kappa_1, p) = \kappa_1^2$ which is unrelated to the nuisance parameter.

Approximate sample sizes required for 80% power of the score tests and the GOF method for various values of the nuisance parameter, with the null and alternative hypotheses are calculated by formulae (11), (12), and (13) and are summarized in

Table 3
Approximate sample sizes required for 80% power of one-sided or two-sided score tests and the chi-square goodness-of-fit test.

κ_0	κ_1	one-sided score test	two-sided score test	GOF method	one-sided score test	two-sided score test	GOF method
			$p = .1$			$p = .2$	
.2	.4	320	360	374	200	242	247
	.6	83	83	94	50	58	62
.4	.6	317	395	404	183	236	241
	.8	73	89	101	42	55	60
.6	.8	232	324	335	132	188	192
	.9	91	136	149	52	80	85
			$p = .3$			$p = .5$	
.2	.4	163	206	208	144	188	188
	.6	40	51	52	35	47	47
.4	.6	142	189	190	121	165	165
	.8	33	46	48	28	41	41
.6	.8	101	146	148	85	126	126
	.9	40	64	66	34	56	56

Table 3. The required sample size is larger for κ_1 close to κ_0 and it is smaller as p approaches 0.5. Sample sizes using the one-sided score test are noticeably smaller than those using two-sided tests. Sample sizes for the two-sided score test are generally smaller than those for the GOF procedure, but when p is close to 0.5, these two sample sizes are similar. As in the power table of Section 3, sample sizes for $p = 0.9$, 0.8 and 0.7 are the same as those for $p = 0.1$, 0.2 and 0.3 in Table 3. Results of sample size comparisons are closely related to those power comparisons.

5. An Example

Twenty pairs of siblings from different families in a community were examined for HIV seropositivity. Of the twenty pairs, two, one and seventeen pairs were classified as both positive, one positive and both negative, respectively: $x_2 = 2$, $x_1 = 1$ and $x_0 = 17$ (HALE and FLEISS, 1993). The positive rate is $\hat{p} = 0.125$ and the estimate of the kappa coefficient, $\hat{\kappa} = 0.77$, suggests a high correlation within a pair of siblings but the standard error, $\mathrm{SE}(\hat{\kappa}) = 0.22$, is relatively large. Considering the small sample size of the study, we can demonstrate that the kappa within a pair of sibling is fairly large, say, $\kappa \geq 0.25$, and test H_0: $\kappa = 0.25$ against H_1: $\kappa > 0.25$ at $\alpha = 0.05$ level. The one-sided score statistic is $z_s = 1.936$ ($p = 0.026$) from (2) and (3) while the two-sided score and GOF statistics are $z_s^2 = 3.75$ ($p = 0.053$) and $X_1^2 = 3.18$ ($p = 0.075$) from (4). The one-sided score test rejects $\kappa_0 = 0.25$ in favor of $\kappa > 0.25$ but the other tests do not. Note that the two-sided score test is more sensitive than the GOF method.

Table 4
Approximate and exact sample sizes required for power $= 1 - \beta$ of tests at $\alpha = 0.05$ when $p = 0.125$, $\kappa_0 = 0.4$ and $\kappa_1 = 0.8$ (the powers in parentheses are exact).

test	$1 - \beta = 0.81$		$1 - \beta = 0.6$	
	approx sample size	exact sample size	approx sample size	exact sample size
one-sided score	61(0.76)	62(0.81)	35(0.61)	34(0.60)
two-sided score	75(0.82)	71(0.80)	47(0.58)	48(0.64)
GOF	84(0.84)	76(0.81)	53(0.59)	55(0.61)

Consider the problem of designing a sibling study with reasonable power. We want to find how many pairs are needed for power = 80% (and 60%) for a test at $\alpha = 0.05$ when $p = 0.125$, with the null and alternative as H_0: $\kappa_0 = 0.4$ and H_1: $\kappa_1 = 0.8$. From (11), (12) and (13), the approximate sample sizes required for 80% power of one-sided, two-sided score and GOF tests are 61, 75 and 84 respectively, see Table 4. Corresponding exact sample sizes are 62, 71 and 76. The approximation for the one-sided score method is virtually the same as the exact while exact sizes for the two-sided score and the GOF methods are 6% and 11% greater than the corresponding exact size. For a 60% power, all three methods provide accurate sample sizes. The saving in sample size using by the one-sided score method is substantial. Required sample sizes using the two-sided score and the GOF tests are 15% and 23% greater required when using the one-sided score test for 80% power, and 41% and 61% large for 60% power.

6. Remarks

NAM (2000) presented an efficient interval estimate of kappa using the score method and showed that the expected length of its interval is shorter than that of a method based on GOF procedure (see, DONNER and ELIASZIW, 1992; HALE and FLEISS, 1993). The advantage of the score method in interval estimation translates to a gain for the score test over the GOF test in power and/or sample size. Interval estimation and the corresponding test are consistent for both the score and GOF methods but not for the usual Wald-type (crude) method. From the data in Section 5, for example, 90% two-sided confidence intervals by the score and Wald-type crude procedure are (0.319, 0.952) and (0.411, 1.132), respectively. The associated significance-tests for the corresponding lower limits of the intervals yield p-values; 0.05 and 0.14 for the score and crude tests, respectively. The former is in agreement with the confidence coefficient while the latter isn't. In addition, the upper limit of the Wald-type crude interval is beyond the permissible bound for kappa. It is important to examine whether the sample size required for a specific power of the score test can also provide the desired precision in the corresponding

interval estimation. Begin a paragraph the exact power requires heavy computation which increases with sample size in a geometrical fashion. Unless the sample size is small, the asymptotic power is reasonably close to the exact and can serve as a good approximation. Intensive computation is involved in searching for the exact minimum sample size for a given power. The sample size obtained by using asymptotic power can be useful as the initial trial value to start searching for the exact sample size, i.e., this initial value is, in general, close to the exact one. The expression for the inbreeding coefficient, F, in genetic studies is identical to that of the intraclass version of the kappa coefficient. In an analysis of phenotypic data, we are interested in whether the Hardy-Weinberg law ($F = 0$) holds, e.g., YASUDA (1968). In studies involving the reliability of ratings, however, we are concerned with reasonably good agreement of ratings and rarely interested in negative or zero kappa. The meaning of a kappa coefficient equal to zero (only chance, or independent) or one (perfect agreement) is unique and not ambiguous but interpretation of intermediate values in terms of degree of agreement is not clear-cut. The specific value of a kappa coefficient for the null hypothesis is related to the nature of the study. In such reliability studies, we intend to show a certain degree of intraclass agreement, i.e., the intraclass correlation coefficient is greater than a desirable minimum value. The one-sided score test is best suited for this purpose. Even when a two-sided procedure is required, the two-sided score method is more powerful than the GOF method. Finally, it is mentioned that the chance-corrected measure, kappa estimate, is always smaller than the concordant rate.

Acknowledgements

The author thanks the referees for helpful comments that have improved the presentation of the paper.

Appendix 1: Asymptotic limit of MLE of p

From Section 2, the asymptotic limit of $\tilde{p}$, $\bar{p}$, is found by solving a cubic equation:

$$a'_0\bar{p}^3 + a'_1\bar{p}^2 + a'_2\bar{p} + a'_3 = 0,$$

where

$$a'_0 = 2(1-\kappa_0)^2,\ a'_1 = -\{3(1-\kappa_0) + p - q\}\ (1-\kappa_0)\,,$$

$$a'_2 = 2p - 2(2 - q^2 - pq\kappa_1)\ \kappa_0 + \kappa_0^2 \text{ and } a'_3 = p\{1 + q(1-\kappa_1)\}\ \kappa_0\,.$$

Letting $b'_i = a'_i/a'_0$ for $i = 1, 2$ and 3, $c'_1 = b'_2 - (b'_1)^2/3$ and $c'_2 = b'_3 - b'_1b'_2/3 + 2(b'_1/3)^3$, we have the asymptotic limit of $\tilde{p}$:

$$\bar{p} = -2(-c'_1/3)^{1/2} \cdot \cos(\pi/3 + \theta'/3)\ - b'_1/3\,,$$

where

$$\cos\theta' = (27)^{1/2} \cdot c_2'/\{2c_1'(-c_1')^{1/2}\}.$$

Appendix 2: An Approximation of Variance of $S_\kappa(\kappa_0, \tilde{p})$ under Alternative

The second-order partial derivatives of the likehilhorsd are

$$\frac{\partial^2 \ln L}{\partial \kappa^2} = -\left\{\frac{x_2 q^2}{(p+q\kappa)^2} + \frac{x_0 p^2}{(q+p\kappa)^2} + \frac{x_1}{(1-\kappa)^2}\right\},$$

$$\frac{\partial^2 \ln L}{\partial \kappa\, \partial p} = -\left\{\frac{x_2}{(p+q\kappa)^2} - \frac{x_0}{(q+p\kappa)^2}\right\},$$

$$\frac{\partial^2 \ln L}{\partial p^2} = -\left(\frac{x_2+x_1}{p_2} + \frac{x_0+x_1}{q^2}\right) - \left\{\frac{x_2}{(p+q\kappa)^2} + \frac{x_0}{(q+p\kappa)^2}\right\}(1-\kappa)^2.$$

Denote the elements of the information matrix are

$$I_{11} = -E(\partial^2 \ln L/\partial \kappa^2),$$
$$I_{12} = -E(\partial^2 \ln L/\partial \kappa\, \partial p) \quad \text{and} \quad I_{22} = -E\{\partial^2 \ln L/\partial p^2\}.$$

Setting $T \equiv (\partial \ln L/\partial \kappa) - (I_{12}/I_{22}) \cdot (\partial \ln L/\partial p)$, we have

$$\text{var}\,(T) = I_{11} - I_{12}^2/I_{22} \tag{A2.1}$$

(e.g., BARTLETT, 1953). A test statisitc, T', is obtained by replacing the nuisance parameter in T by a consistent estimator of p for a given $\kappa = \kappa_0$. T' and T are stochastically equivalent in large samples, and $\text{var}\,(T') = \text{var}\,(T)$ in probability under $\kappa = \kappa_0$. Note that $T' = S_\kappa(\kappa_0, \tilde{p})$ when a consistent estimator of p in T' is the MLE of p for a given $\kappa = \kappa_0$.

Consider the partials evaluated at $p = \tilde{p}$ and $\kappa = \kappa_0$, i.e.,

$$\left(\frac{\partial^2 \ln L}{\partial \kappa^2}\right)_{p=\tilde{p},\,\kappa=k_0} = -\left\{\frac{x_2 \tilde{q}^2}{(\tilde{p}+\tilde{q}\kappa_0)^2} + \frac{x_0 \tilde{p}^2}{(\tilde{q}+\tilde{p}\kappa_0)^2} + \frac{x_1}{(1-\kappa_0)^2}\right\},$$

$$\left(\frac{\partial^2 \ln L}{\partial \kappa\, \partial p}\right)_{p=\tilde{p},\,\kappa=k_0} = -\left\{\frac{x_2}{(\tilde{p}+\tilde{q}\kappa_0)^2} - \frac{x_0}{(\tilde{q}+\tilde{p}\kappa_0)^2}\right\}, \tag{A2.2}$$

$$\left(\frac{\partial^2 \ln L}{\partial p^2}\right)_{p=\tilde{p},\,\kappa=k_0} = -\left(\frac{x_2+x_1}{\tilde{p}^2} + \frac{x_0+x_1}{\tilde{q}^2}\right) - \left\{\frac{x_2}{(\tilde{p}+\tilde{q}\kappa_0)^2} + \frac{x_0}{(\tilde{q}+\tilde{p}\kappa_0)^2}\right\}(1-\kappa_0)^2.$$

Recall that $\bar{p}$ and $\bar{q}$ are asymptotic limits of $\tilde{p}$ and $\tilde{q}$ (Section 3.1, Appendix 1). The expectation of the negative of (A2.2) under H_1: $\kappa = \kappa_1$ as $\tilde{p}$ and $\tilde{q}$ approach to $\bar{p}$ and $\bar{q}$ in the limit converge in probability to

$$\begin{aligned} I^*_{11} &\equiv -E_1\{(\partial^2 \ln L/\partial\kappa^2)_{p=\tilde{p},\,\kappa=\kappa_0}\} = n(\bar{q}^2\phi_1 + \bar{p}^2\phi_2 + 2pq\phi_3)\,, \\ I^*_{12} &\equiv -E_1\{(\partial^2 \ln L/\partial\kappa\,\partial p)_{p=\tilde{p},\,\kappa=\kappa_0}\} = n(\phi_1 - \phi_2)\,, \\ I^*_{22} &\equiv -E_1\{(\partial^2 \ln L/\partial p^2)_{p=\tilde{p},\,\kappa=\kappa_0}\} = n\{f_1\phi_1/\bar{p}^2 + f_2\phi_2/\bar{q}^2 + f_3/(\bar{p}\bar{q})^2\}\,, \end{aligned} \tag{A2.3}$$

where

$$\begin{aligned} &\phi_1 = p(p + q\kappa_1)/(\bar{p} + \bar{q}\kappa_0)^2\,, \qquad \phi_2 = q(q + p\kappa_1)/(\bar{q} + \bar{p}\kappa_0)^2\,, \\ &\phi_3 = (1 - \kappa_1)/(1 - \kappa_0)^2\,, \qquad f_1 = 2\bar{p}(\bar{p} + \bar{q}\kappa_0)\,(1 - \kappa_0) + \kappa_0^2\,, \\ &f_2 = 2\bar{q}(\bar{q} + \bar{p}\kappa_0)\,(1 - \kappa_0) + \kappa_0^2 \quad \text{and} \quad f_3 = 2pq(1 - 2\bar{p}\bar{q})\,(1 - \kappa_1)\,. \end{aligned}$$

For κ_1 in a neighborhood of κ_0, we may approximate the asymptotic variance of T' under H_1: $\kappa = \kappa_1$ as $\mathrm{var}_1\{S_\kappa(\kappa_0, \tilde{p})\} \approx I^*_{11} - I^{*2}_{12}/I^*_{22}$ from (A2.1) and (A2.3). A result of simulations showed that a value of the approximated variance is satisfactorily close to an empirical value when a difference between κ_0 and κ_1 isn't great.

References

BARTLETT, M. S., 1953: Approximate confidence interval, II. More than one unknown parameter. *Biometrika* **40**, 306–317.

BLOCH, D. A. and KRAMER, H. C., 1989: 2 × 2 kappa coefficients: measure of agreement or association. *Biometrics* **45**, 269–287.

COHEN, J., 1960: A coefficient of agreement for nominal values. Educational and Psychological Measurement **20**, 37–46.

DONNER, A. and ELIASZIW, M., 1992: A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. *Statistics in Medicine* **11**, 1511–1519.

DUNN, G., 1989: Design and Analysis of Reliability Studies. New York: Oxford University Press.

HALE, C. A. and FLEISS, J. L., 1993: Interval estimation under two study designs for kappa with binary classifications. *Biometrics* **49**, 523–534.

HAYNAM, G. E., GOVINDARAJULU, Z., and LEONE, G. C., 1962: Tables of the cumulative noncentral chi-square distribution, Case Statistical Laboratory, Publication No. 104, Part of the tables have been published in Selected Tables in Mathematical Statistics Vol. 1. Edited by HARTER, H. L. and OWEN, D. B., Markham, Chicago.

MAK, T. K., 1988: Analyzing intraclass correlation for dichotomous variables. *Applied Statistics* **37**, 344–352.

MENG, R. C. and CHAPMAN, D. G., 1966: The power of chi-square tests for contingency tables. *Journal of the American Statistical Association* **29**, 965–975.

MITRA, S. K., 1958: On the limiting power function of the frequency chis-quare test. *The Annals of Mathematical Statistics* **29**, 1221–1233.

NAM, J., 1995: Sample size determination in stratified trails to establish the equivalence of two treatments. *Statistics in Medicine* **14**, 2037–2049.

NAM, J., 2000: Interval estimation of the kappa coefficient with binary classification and an equal marginal probability model. *Biometrics* **56**, 583–585.

SCOTT, W. A., 1955: Reliability of content analysis; the case of nominal scale coding. *Public Opinion Quarterly* **19**, 321–325.

SMITH, C. A. B., 1970: A note on testing the Hardy-Weinberg Law. *Annals of Human Genetics* **33**, 377–383.

USPENSKY, J. V., 1948: Theory of Equations. New York; McGraw-Hill.

WRIGHT, S., 1951: The genetical structure of populations. *Annals of Eugenics* **15**, 322–354.

YASUDA, N., 1968: Estimation of the inbreeding coefficient from phenotype frequencies by a method of maximum likelihood scoring. *Biometrics* **24**, 915–935.

JUN-MO NAM
Biostatistics Branch
DCEG
National Cancer Institute
Executive Plaza South
Room 8028
6120 Executive Boulevard
MSC 7240
Rockville, Maryland 20892-7240
USA
e-mail: namj@mail.nih.gov

Received, January 2000
Revised, February 2001
Revised, August 2001
Accepted, January 2002